DE GRUYTER OPEN

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **17**, No 2
Sofia • 2017 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.1515/cait-2017-0013

Review on Big Data & Analytics – Concepts, Philosophy, Process and Applications

Kari Venkatram, Geetha Mary A.

School of Computing Science and Engineering, VIT University, Vellore 632 014, Tamil Nadu, India
E-mails: venkat_kari@yahoo.com geethamary@vit.ac.in

Abstract**:** *Big Data analytics has been the main focus in all the industries today. It is not overstating that if an enterprise is not using Big Data analytics, it will be a stray and incompetent in their businesses against their Big Data enabled competitors. Big Data analytics enables business to take proactive measure and create a competitive edge in their industry by highlighting the business insights from the past data and trends. The main aim of this review article is to quickly view the cutting-edge and state of art work being done in Big Data analytics area by different industries. Since there is an overwhelming interest from many of the academicians, researchers and practitioners, this review would quickly refresh and emphasize on how Big Data analytics can be adopted with available technologies, frameworks, methods and models to exploit the value of Big Data analytics.*

Keywords**:** *Data, Big Data, Hadoop, no sql, analytics.*

1. Introduction

Data manifestation has been changed radically over last few years. Today's data is much different from the past. Data is becoming amorphous, which means many forms of data without any proper shape or form. Big Data solutions came forward to handle such types of data. Big Data is referred as Data Intensive Technology (DIT)[1]. There are significant learnings from the continuous data generation from both science & technology as well as quality of the business and lives [2]. Big Data is becoming part of every aspect of our lives. Technologies such as hadoop, cloud computing and visualization enable data collection, process & store and visualize data insights [1]. Google search was built based on Big Data considerations and google analytics helped in arresting the virus spread H1N1 in the USA during virus spread. Big Data and its analytics will be every part of our life in future, right from the schooling, i.e., identifying the right school for your kids, helping human to monitor their health suggesting the remedies for wellbeing and it may recommend you for right partner based on your interest and many more.

2. Evolution of Big Data technology

From the ancient times human knows how to store and process data, the initial storage being on stones of caves and tamarapatra (leaves), etc., which was referenced by next generations to interpret the previous generation's wisdom and heritage. After invention of Paper around 220AD the data was stored in bunch of papers, which was nothing but a book. However, the ability to analyse the data is limited to one's own brain and knowledge inferred by an individual, which was stepping stone for inventing an alternate for human brain that is computer evolution. This era began with punch cards and OMR for data storage.

Since the computer became integral part of every business there was a need to communicate among computers in many geographical locations, which led to invention of Internet, which triggered massive improvement in data storage, processing and analysing techniques that started digital electronics era. In order to manage and process the data, files systems and database management systems have been developed. File systems enable to store and access hard disk and DBMS enables to store and access data. Data generation rate has been tremendously increased from last two decades and it is the origin for new horizons in data storing and processing requirements for wide variety, huge volumes of data from kilobytes to zeta bytes.

2.1. Data transformation to wisdom

Information, knowledge and wisdom extracted from the data. *Data* [3, 4] is nothing but gathering some of the facts, specifications, objectives, particulars, some details, statistics, figures, images, audio and video, etc. *Information* [3, 4] is something derived after analysing particular data that is being used for taking some of the business decisions. It is derived by synthesizing and contextualizing the data to provide value. *Knowledge* [3, 4] is something we get more understanding and expertize from the information on some subject. *Wisdom* [4] is a technique to utilize particular knowledge at some situation. This is also termed as solution pattern for a particular problem.

2.2. Data storage evolution

Data storage and retrieval is being controlled by the operating system using file operations in the underlying file system. Different operating systems follow different approaches to manage their files in the storage devices. Later relational data base management systems evolved to address the redundancy issues in the traditional database management systems by means of table. RDBMS has been widely used in many of the business applications for data storage and processing. It has been a widespread across the enterprises. However, it has a limitation to work with heterogeneous systems as it can support only limited data types and hence there is a need for custom data types. To overcome this issue a new data base management systems called object relational database management system has been invented. It enables to define custom / own data types and methods. ORDBMS support data types such audios, videos, images and any custom data structures.

2.3. Attributes of data in five C's [3]

As described in the Fig. 1, data has five important attributes. We call them simply 5 Cs such as Clean, Consistent, Confirmed, Current and Comprehensive. All the attributes are briefly defined in the following paragraph.

Clean: Data in raw form requires cleansing to get into a desired format to gain some information from it. *Consistent:* Data to be represented in a way that is consistent in its representation, so management can take a decision without having any arguments among the teams. *Confirmed:* Data to be confirmed to adapted dimensions across the data and should be consistent to be utilized by any team. *Current:* Data should be accurate at that point of time to take some business decisions. *Comprehensive:* Data should be inclusive of all the required items from its granular level from all the sources to take some decision.

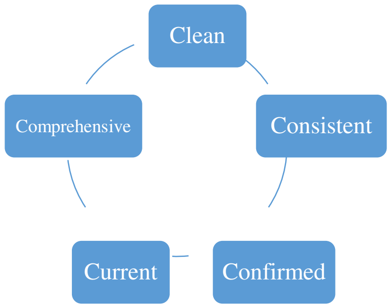

Fig. 1. Attributes of data

2.4. Data processing evolution

Storing and processing of large data is a challenge, there are some innovations done to address these technical and processing challenges. Ex: Grid computing handles voluminous data, cloud computing deals with high velocity and huge data volumes, open sourcing technologies for cost reduction and virtualization reduced time to test, deploy and improves processing speeds, etc. However, there are issues with these solutions like grids are expensive, clouds are seeming to be slow, open source less robust, virtualization tends to slowdown the execution process. So, there is a need for new solution to handle the data challenges.

As depicted in Fig. 2, ERP generates limited data in some megabytes through a form of purchases and payments details, etc. CRM generates more data in some of gigabytes and process data about customers, relationships, offers and other details like segmentation etc. ERP and CRM applications data managed using OLTP. With the evolution of web technologies lots of data shared among a web servers, hence data generation increased significantly through web logs. At the same time data analytics requires historical information to extract data insights. Hence many of terabytes data were generated. Also, there is a new requirement for storing the historical data in OLAPs apart from OLTP, which manages transactional data. In the last few years data has exponentially grown due to IOT (Internet Of Things) [5], mobile apps, sensor data, etc., which requires petabytes of data. Hence there is a need of distributed file system such as Hadoop to maintain the Big Data.

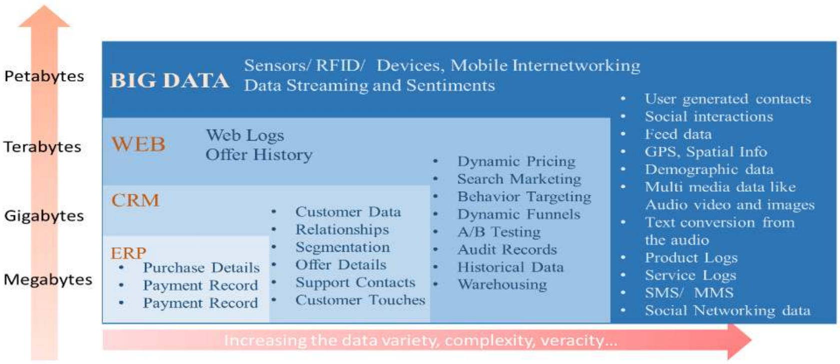

Fig. 2. Data growth through technological evolution

OnLine Transaction Processing (OLTP) systems or operational systems are relation database management systems [4], process a huge number and short on-line transactions. It maintain data consistency with low redundancy real time data processing or data streaming [5].

OnLine Analytical Processing (OLAP) systems follows write once and read more paradigm. It uses very complex queries [4] with many aggregative functions in the query. It is used in data mining and analytics. Data Cube and OLAP are two popular models in early analytics [6]. Generally data processing done in long running data batch jobs, it is called as batch processing [5] Jobs.

Brief history of data evolution with milestones [7]: Megabyte to Gigabyte (1970-80), database machines generate data in GBs. Gigabytes to Terabytes (1980-‘00), shared nothing parallel databases to generate data in TBs. Terabytes to petabytes (2000-10), GFS and Map Reduce to generate data in PBs. Petabytes to Exabytes (2010 +), IoT, Cloud, Big Data enablers, etc., to generate data in EBs

2.5. Evolution of data warehouse [4]

In the initial days, OLTP systems were used for analytics to get the data insights, though it is advantages to get insight of real time data but analytical functions with multiple joins on very large number of records is a performance hit. Having both operational and analytical functions working on the same data base is a challenge as usability and performance considerations. In order to make quick decisions from the data, organizations started adapting Decision Support System (DSS) with help of tools, algorithms and techniques to get data insights. The data aggregation is done in the analytical models of the data warehouse and ETL (Extract Transform Load) has been developed to extract the data, transform the data and load them into the data warehouse form the operational databases.

Data WareHouse (DWH) [4] is a decision support system built as a data driven system, enables management to take decisions. Usually it focusses on single subject of the data sometimes it termed as Data mart.

Enterprise Data Warehouse (EDW) [4] is a system based on OLAP model, intended for analyzing the data and generating the insights out of the entire enterprise data as form of Reports. EDW will focus on all the subject/ functional areas of the entire business and define the business rules to derive the insights of the data across the functions, such as Customer, Product, Sales, Services, Finance, Marketing, Manufacturing, etc.

2.6. Evolution could computing

Storing the data into the database servers, managing the data, maintaining data centres and the underlying infrastructure is cumbersome for non IT organizations. Organizations have to invest a lot on the infrastructure and intellectual property for maintaining the same. Many of the organizations would like to focus on their core business rather than maintaining these data centres and required infrastructure, as they focus to gain profits by just focusing on their business strategies. The idea of providing all these services for managing infrastructure, platforms and software have been evolved recently. To manage these services with lots of elasticity and on demand based services, a provider requires to maintain the infrastructure with distributed computing power.

Grid Data Computing (GDC) [8] is achieved by set of computers and resources work for some common requirement [9]. It works in basis of distributed architecture and non-interactive workloads.

Distributed Data Computing (DDC) is something that usually manage a pool of computers to work together in by message passing technique to achieve the required goal. Parallel execution of tasks and consolidation at the end is done before sending the response.

Cloud Data computing (CDC) [8]: Since data management and data processing is being increased day by day, it became difficult to individual companies to manage the data bases and data centres. Based on the demand companies need to increase the data processing and management capabilities and it is the main factor for evolution of Cloud computing [9]. The idea of cloud computing is to provide all the required infrastructure as a service on demand within a short time. This will greatly reduce the investment on an infrastructure by an individual company. Third party data centres are the hub for Cloud computing operations such as data storage and solutions provided to users and enterprises. Cloud computing is a shared infrastructure and services among the cloud users, hence it will be most economical way for data storage and processing. Undoubtedly, Cloud computing is one of the technology swifts in this century [5]. Three Characteristics of cloud computing are: 1) Pay per usage; 2) Flexible, and 3) Infrastructure Management [5].

To store, manage and process the data on the cloud it uses terrestrial servers across the internet. Cloud is a distributed parallel computing mechanism works in-line with Grid computing technology. These servers collectively process the data. Cloud computing is synonymous to Service Oriented Computing (SOC). Also, all the resources such as software, platform and hardware are delivered as a service. It is provisioned over the internet or other private networks. Three most popular cloud computing components are Software as a Service (SaaS) [5], Platform as a Service (PaaS) [5], Infrastructure as a Service (IaaS) [5].

Cloud deployment models: Cloud services can be deployed in the four models such as: *Private Clouds* [5] – a model where cloud infrastructure is being operated and managed solely for a single organization. *Public Clouds* [5] – a model where the services are available on the internet and is open for public; concern is the security in the public clouds as it is a shared pool for public. *Hybrid Cloud* [5] – a combination of both private and public cloud. *Community Clouds* – a model for like-minded organizations with the common goals share the infrastructure, services, etc., among them.

2.7. Modeling towards Big Data

There are many definitions for Big Data, we may refer to huge data sets where the size of this referred data is not in a position to manage by any traditional data base management systems for capturing, storing, analysing and managing the datasets. We come across data in every possible form, whether it is through social media sites or sensor networks or digital images or any videos or cell phone GPS signals or purchase transaction records or web logs or individual medical records or any archives or may be military surveillance or e-commerce data, complex scientific research data or any other form. Altogether it amounts to over some large data! This data is what we may call as…BIG DATA!

3. Big Data – A deep dive

As discussed in the previous section, Big Data is something like a set of huge data sets which are complex and requires tedious jobs to capture, store, process and analyse them. The definition of "BIG DATA" may vary from organization to organization, person to person depending upon their use cases and their value generation from their data and data characteristics such as data size, capacity, competence of human resource, techniques used for analysis and, etc. For example, some organization, managing few GB of data may be a cumbersome job where as for others it may be some terabytes. Big Data may be referred to data which is being generated in a very large quantity / volume at a high velocity/ rate in many different formats of data. As we store the data, we have to plan to analyse the data and get the insights of it and take some decision based on the insights what we get from the data. However, as a traditional practice only some of the sample data taken for analysis instead of taking complete data due to many technical challenges to handle complete data set. Data was sampled and analysis was done on the sample data for decision making. However, with the help of Big Data and associated technologies and frameworks such as Hadoop, we would be able to process and analyse complete data set. So, we can achieve very accurate results from the complete data set as it is not biased for decision making.

Data is rapidly growing and to quantify the data units also increased day by day. As mentioned in Table 1, the different units and its decimal and binary equivalent values shown. Yottabyte is the largest unit with a decimal value of 10 power 24 and binary value of 2 power of 80.

Table 1. Data volume

Name	Decimal value	Binary value		Name	Decimal value	Binary value
Kilobytes (KB)	10^3	2^{10}		Petabyte (PB)	10^{15}	2^{50}
Megabytes (MB)	10^6	2^{20}		Exabyte (EB)	10^{18}	2^{60}
Gigabytes (GB)	10^9	2^{30}		Zettabyte (ZB)	10^{21}	2^{70}
Terabytes (TB)	10^{12}	2^{40}		Yottabyte (YB)	10^{24}	2^{80}

In a study IDC digital universe has predicted that the growth of the data would go up to 40K Exabytes that is the 50 fold growth of the data by 2020 [10]

Digital data generation starts from mainframe computers, the data generation speed and volume are very limited. There were few thousands of users and few hundreds of applications at that time, later client server technology comes into existence and data speed and volume of the data also increased. With client server technology, thousands of applications come into existence and millions of users are using these applications. With an internet and web technologies multiples of ten thousand of applications are being used by hundreds of millions of users. With internet and mobile communication come into lime light, lots of data generated with mobile technologies, social media, sensors, clouds, etc., coming into this generation we are rapidly generating very huge data sets in different formats.

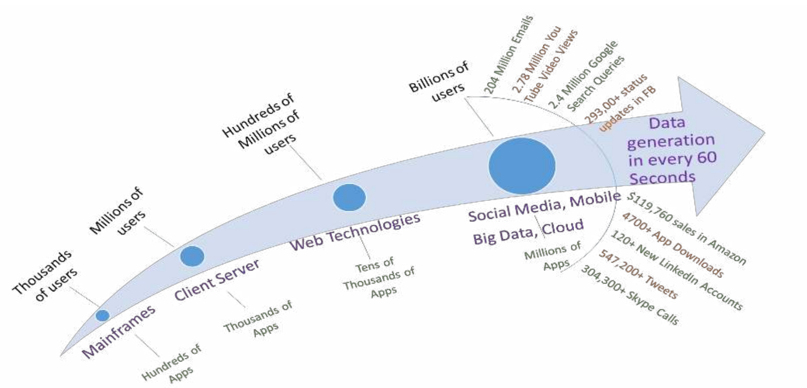

Fig. 3. Data transformation from main frames to Big Data and data generation every 60 seconds

As mentioned in Fig. 3, 204 million emails are sent every 60 seconds, 2.78 million of status updates are happening every minute [11]. Consider 90% of the data in the world now is generated in the last two years. This enormous data has left the organization to grapple how to make use of this data.

3.1. Need for Big Data

Big Data is becoming very important in our day to day life. It is also becoming really critical to us as it is not just emerging as one of the most powerful technologies in this era, also it is helping us to take day to day decisions. Big Data enables to analyse the complete data set instead of some sample data, hence there is non-biased outcomes. Social media networking enables analytics in a very cost effective manner and quick feedback and reviews from customers. It is possible to build a product based on customer preferences with the help of Big Data. In the healthcare centres, doctor can provide the solution referring to the previous history of the patients. Big Data getting Normalized [12]. Here *Normalization* is referred as a process that enables to generate ideas and actions from the Big Data and it is taken as just a normal process as day to day life. This becomes a standard practice and part of our lives. This may be used to disclose the habits of one's eating, acting and thinking based in a situation and provides appropriate suggestions.

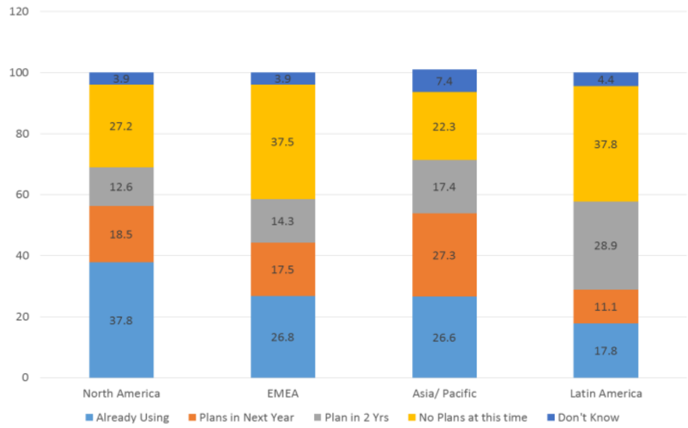

Fig. 4. Big Data spread

Big Data Presence: As shown in Fig. 4, Gartner has done a survey [11] and described the presence of Big Data across the world. North America region top among the other regions 37.8% are adapted Big Data and 18.5% are ready to adapt Big Data. Latin America being the down side among the others adapted 17.8% and planned to adapt 11.1% soon.

3.2. Characteristics of Big Data

As depicted in the Fig. 5, Big Data has defined by the below seven characteristics [13, 14] such as Volume, Velocity, Variety, Veracity, Validity, Volatility and Value.

Volume [1, 15, 16]: Big Data means to someone as enormous volume of data. As we have discussed Big Data deals very large datasets, it is referred to large volume. For an instance, 15 terabytes of posts on Facebook could mean Big Data!

Variety refers to data in different forms or types [1, 15, 16]. Data such as emails, videos, pdf's, logs, images, excel sheets, etc., are different varieties of data. Any data can be organized into one of the below mentioned three types [5] such as structured data, semi-structured data and unstructured data.

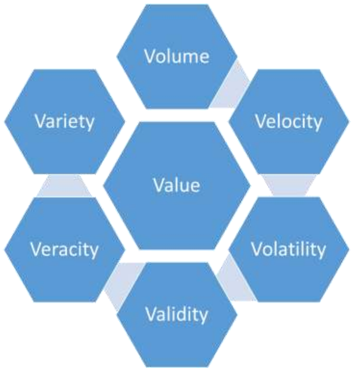

Fig. 5. Characteristics of Big Data

Velocity is pace at which data generated and flowing from the data sources [1, 15, 16] like business process, machine generated data and human interactions with computer, etc. In simple data generated by machines, processes, human interactions, sensor generated data are generated at a high velocity. Velocity means the rate at which data is flowing in to the systems. So, Big Data requires fast processing to deal with the velocity of the data. Time for processing this data plays a very crucial role as we need to take right decision at the right time in any of the organizations.

Veracity refers to the biased such as partial, falsity such as inaccuracy, noisy [1, 16], such as unclear and abnormalities in data. It is very challenging job to deal with veracity in data for doing analysis. It requires to clean and process the data make the data certainty and truth and consistent. Veracity refers to the uncertainty of the data, that is the data can sometimes get messed up and it is very difficult to interpret and trust the data. Due to uncertainty of data, trust worthiness comes in to play and it is not considered for decision making.

Validity is like veracity, where data has anomalies, other issue of the Big Data is validity of the data at that moment. This means the data what we use is correct and accurate for the intended use at the given point. Hence the valid data is key for making the right decisions.

Volatility in Big Data, may be referred to how long is the data valid and how long must it be stored in our systems. Organizations would need to plan and determine at what point the data is no longer relevant for the data analysis.

Value refers to the processing of the data that produced out after analysis [1]. The value of data is not for just one time use it can also be reused for any future purpose analysis by combining with other data sets.

3.3. Big Data databases

Today lots of data generated with personal information, social graphs, geographic location, user-generated data though social and mobile media communications and tools such as Twitter, Facebook, WhatsApp and Google+, etc. Data is exponentially increasing on day by day with all social networking data, sensor data, multimedia data and enterprise data. To avail the benefit of the Big Data properly, it is required to store and process huge number of datasets. These kinds of datasets can't be handled by traditional SQL databases, as these were never designed to cater these kinds data.

Hence the new generation of No SQL databases are evolved, which can handle these huge datasets, high velocity data and wide varieties properly.

Data lakes are repositories, where set of multiple data storage instances from several data assets in an organization [5] are stored. The major role of a data lake is to ensure all the data in a crude form from all the data sources of an organization. Simply it is a dump of all the data into one cluster. Data lakes can store any data types. It is a schema less approach and it tuned for data retrieval in a faster way. Since load first paradigm [17] enable an enterprise to consume much data from multiple sources and technologies, this is being enabled through data lakes.

No SQL to be read as Not only SQL [18], is a non-relational database management systems [19]. These databases are different from traditional relational database management systems in some ways. No SQL is designed for storing huge data sets in distributed environment. These type of data bases may not require to have fixed schema, it avoid join operations and typically it can scale out approach. Nowadays lots of unstructured data generated and we need to have some way to store them into databases. No SQL is the solution for storing such types of data. It is clear that, in the future NoSQL databases going to be more widely used to handle all such data as schema-less database would be more pronounced [19]. Some of the prominent and widely used NoSQL Databases are MongoDB, Redis, DataStax, MarkLogic and DynamoDB, etc.

The comparison of RDBMS with NoSQL database is shown in Table 2. Advantages and flexibilities of No SQL databases are mentioned in the below table.

Table 2. Comparison between RDBMS and No SQL

RDBMS	No SQL
Structured and organized data: Data stored in columns and rows	Stands for Not Only SQL, it can store any structured, semi structured and unstructured data
Structured Query language (SQL)	No declarative query language
Data and its relationships are stored in separate tables. Pre-defined schema	No predefined schema
Data Manipulation Language, Data Definition Language	Key-Value pair storage, Column Store, Document Store, Graph databases
Tight Consistency	Eventual consistency rather ACID property
BASE Transaction	Unstructured and unpredictable data
	Based on CAP Theorem
	Prioritizes high performance, high availability and scalability

3.3.1. CAP Theorem

No SQL database is a distributed architecture follows CAP theorem, built on three characteristics depicted in Fig. 6, such as:

Consistency: This means that the data in the database remains consistent after the execution of an operation.

Availability: This means that the system is always available, which means no downtime.

Partition Tolerance: This means that the system continues to function even the communication among the servers is unreliable.

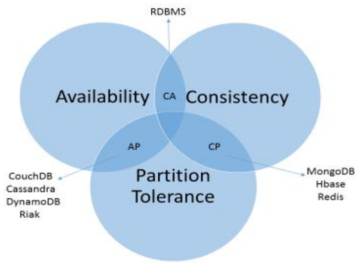

Fig. 6. CAP Theorem

In Practical it is impossible to fulfil all three requirements stated above. CAP provides the basic characteristics for a distributed system to follow two of the three characteristics. Therefore, all the current NoSQL database follow the different combinations of the C, A, P from the CAP Theorem. The advantages and disadvantages of No SQL are shown in the Table 3.

Table 3. No SQL advantages and disadvantages

Pros	Cons
Highly scalable as it Scale out horizontally	No proper standardization
Distributed Computing	Limited query capabilities
Lower cost as it can work on commodity hardware and scale out approach	Joins are not possible
Flexible Schema adoption for semi-structure data	
No complicated Relationships maintained	

3.3.2. Categories of No SQL

NoSQL data bases can be categorized into four general types. Each of the types has its own pros and cons based on their specifications. Based on the requirement need to choose the best type suits our requirement. We can't distinguish one is superior to other type. There are: 1) Key-Value stores; 2) Column-oriented databases; 3) Graph data base; 4) Document oriented data base.

Key-Value store: It is one of the basic types of No SQL data store [5, 7]. It is designed to manage large datasets. It was built on the specifications provided in Amazon's Dynamo[5] white paper. Since it is key value stores, it enables schema-less data storage capability. In this, data is represented as hash table. The values in the table are stored as JSON, String or BLOB, etc. The key of the hash table usually string or any other object such as hashes, list, set, etc. It will have key and value pairs of the data being stored for each row. Key-Value stores Database mostly follows the 'Availability' and 'Partition' aspects in CAP theorem. This kind of Key-Values stores suitable for shopping cart contents, etc. Example of Key-value store Data Base: Redis, Dynamo [5], Riak, etc.

Column oriented databases primarily work on columns unlike RDBMS databases [5, 7]. Each column is here treated independently. In this all the data for a column are stored contiguously. Column specific files for each column is maintained. Query in column oriented DB is built and work on columns. Since the data with in a column will have the same type of data it is easy for comparison within a data file. It gives better performance in querying the details it can directly access specific column data from the column data file. Row stores data in a row format. It is easy to add or modify a record, however we need to fetch some unnecessary data to read. Row-stores are used usually see in RDBMS databases. Whereas column stores used for write once and read many databases. It is suitable for read intensive large data repositories such as data warehouses. Hence the column aggregation functions such as count, sum, average, minimum and maximum will have a greater performance experience with column oriented databases. Majorly these databases are seen in data warehouses for building business intelligence from Customer Relationship Management (CRM), Library card catalogues, etc. Example of Column-oriented databases: BigTable, Cassandra, SimpleDB, etc.

Graph data base is a hybrid data base [5], which can be backing unusual types of data store such as graph in data stores [7]. This structure follows set of ordered pairs called as edge and arcs for each entity. This entity is called as vertices or node. It has a capability to represent any kind of data elegantly with high accessibility. It is a collection of nodes and edges; each entity is represented by a node or vertices and every connection is relationship represented by an edge in the graph between two nodes or vertices. It is uniquely identified by each node and its edge definition. All the adjacent nodes will be known by the neighbour nodes. It uses indexing for look ups. Though the number of nodes increases no of hops remains same. It is an easier way to represent any hierarchal data in a graphical representation and easy to denote and fetch the hierarchal information from the graph databases. Example of Graph databases: OrientDB, Neo4J, Titan, etc.

Document oriented data base is a collection of documents [5]. Data model is represented in the document itself. It is a collection of key value pairs in a document. Key allows to access the value from the document. It follows flexible schema and ease of change in schema in document oriented data bases [7]. In order to group the different kinds of data, documents are stored into collections. Document maintain data in a different key value pairs or key and array pairs. One of the important features of the document oriented database is, it offers nested documents as well, which means essentially, we can have a document as a value for one key in the document. This nested document is generally used for representing the hierarchal data or complex data sets. The comparison between document model and RDBMS is shown in Table 4. Collection in the document model represents table. Each document is like a row of a table. Column and its values are represented by key value pairs. Generally, document model will not support joins.

Table 4. Comparison of RDBMS with document model

RDBMS	Document Model
Tables	Collections
Rows	Documents
Columns	Key/value pairs
Joins	Not available

Examples of Document Oriented databases are MongoDB, CouchDB, etc.

3.4. Big Data usage across industries

Most of the organization across the sectors are exploiting the Big Data benefits. It is aggressively used in certain sectors and they are getting benefited out of Big Data and analytics. Majorly banking, government, media and communication, manufacturing, retailers, e-commerce and social media sectors are largely using the Big Data. Now health care providers, Insurance companies and transportation companies are emerging in the Big Data analytics usage.

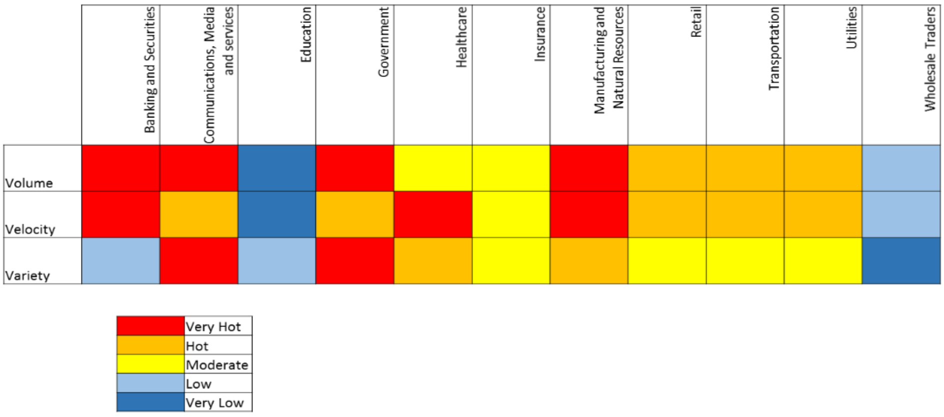

Fig. 7. Big Data – Heat map

Fig. 7, compares the Big Data adoption by different industries [20]. Banking, securities, communication and media, manufacturing and natural resources are the majorly adopting Big Data analytics. Big Data usage is very hot in these industries. There are many use cases exists in these industries and they are already enjoying the benefits of Big Data. Industries such as government sectors, retail, transportation and utilities are also using Big Data heavily. Education and wholesale traders' industries are having low presence in the Big Data area.

3.5. Analytical data

Organizations using analytics for taking proper decisions to improving their business. Organizations can analyse the data and identify customer buying patterns, customer churn outs, profitability and performance characteristics of their suppliers such as over time, better supply chain management to classify them. Data analytics enables to identify the failure patterns of a product based on product behaviour data by scrutinizing the product behaviour for some period of time. All these data to be stored in an enterprise data warehouse with a support of data marts and with the support of aggregation functions and algorithms. Enterprise data warehouse will have a large number of fact tables surrounded by some of the key dimensions such as customer, supplier, account, location, product and partner, etc., for building their analytics and business intelligence.

4. Big Data analytics

4.1. Data analytics

Data Analytics (DA) [21] is a science, which will examine the data and draws insights out of the information from the data. Eventually it will enable us to monitor the data and make us understand what has occurred, why is it occurred, what is going to occur? and what is to be done to avoid this occurrence?, etc. Hence data analytics is used by industries to make better business decisions based on the historical data using some data analytics tools. Data analytics is renowned from data mining by its scope, focus and the purpose of the data analysis. Data Mining is the process to discover the patterns in the data and establishing unknown relationships among the data. Data analytics extrapolates the conclusion based the knowledge of the person or system with a prior knowledge. Data analytics is used to derive insights of the data starting from OnLine Analytical Processing (OLAP) to CRM analytics. For an example, banks will analyse the data of their customers on their usage patterns of money withdrawals, spending habits, etc., to prevent any fraud in transactions. Another example on any e-commerce site, they would like to understand the buying patterns of the consumers in the site and offer recommended products and provide some discounts to attract the consumers

4.2. History of Big Data analytics

The Big Data concept has been around for some years. Now many of the organizations understand the importance of Big Data, so they started capturing all the data that flood into their businesses. These huge data can be applied for analytics and get significant value from it. Interestingly even before few decades anyone uttered the term "Big Data," many organizations used some basic analytics using spreadsheets, examined the data and uncovered lots of insights and trends from the data captured in the spreadsheets. Till few years ago companies used to gather information, run the analytics on the data gathered and reveal the information used for future decisions. However today organizations are identifying insights for immediate decisions on the whole data they have been using the Big Data capabilities. Abilities like quick action, stay agile will boost organizations into a competitive edge.

4.3. Why is Big Data analytics important?

Big Data analytics is the way to help organizations harness their data to identify new opportunities. That, in turn, the Big Data analytics insights and decision may leads to smarter moves for business growth, improves day to day business operations, manages to get higher profits and happier customers and improves the customer satisfaction. As per Tom Davenport, IIA Director of Research, mentioned in his research article that many of the companies got value in the following ways with Big Data analytics as shown in Fig. 8.

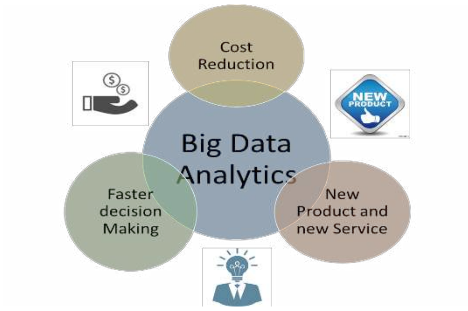

Fig. 8. Big Data analytics – value

Cost reduction: Hadoop and Cloud based analytics are Big Data analytic platform and technologies, which can bring in significant cost advantages especially for storing large amounts of data and process. Also by performing analytics on the Big Data they can clearly identify the ways for doing efficient business.

Faster: Better decision making: For a faster processing of data and analysis, new sources of data, technologies like Hadoop and in-memory analytics companies are able to analyze information real-time so as the decisions are taken at the right time based on what they've analyzed.

New products and services: Through Big Data analytics organizations gain ability to gauge customer needs and understand the sati factionary requirement. With this in mind it enables the organization to think what their customer wants. Many companies are manufacturing new products to meet customers' needs based on the analytics.

4.4. Analytics on Big Data [15]

Just having Big Data alone is useless, you get some insights out of them which is required for taking any business decision to the growth of the enterprise[22]. We need tools and frameworks to highlight the insights from huge datasets, rapidly increasing data and a large variety of data of an enterprise. In the subsequent sections of this document, discussed on the methods and tools used for analysing Big Data analytics in the industry.

Big Data analytics have been embraced as a disruptive technology that will reshape business intelligence which is a domain that relies on data analytics to gain business insights for better decision-making. There is a need of state-of-the-art Big Data analytics tools which are much innovative, highly scalable and powerful for analyzing into insights with multidimensional data visualization [5]. There are many tools built by many people for Big Data visual analytics. Few of them are open source and some of them are commercial. For an instance IBM's InfoSphere and BigInsights [5] are commercial licensed tools for Visual analytics.

4.5. Types of Big Data analytics

There are four types of Big Data analytics: *Prescriptive Analytics* – This type of analysis reveals what actions should be taken; this is the most valuable kind of analysis and usually results in rules and recommendations for next steps. *Predictive*

Analytics – An analysis of likely scenarios of what might happen; this kind of analytics are usually a predictive forecasting. *Diagnostic Analytics* – A look at past performance to determine what happened and why it happened. *Descriptive Analytics* – What is happening now based on incoming data. To mine the analytics, you typically use a real-time dashboard and/or email reports; this is like a report, dashboards, score cards, etc.

4.6. Techniques and methods for Big Data analysis [23]

Based on statistical and mathematical models, some computerized techniques have been developed, which can be used to analyse datasets and provide some insights of the data as a result. There are numerous techniques available in Big Data analytics. All these methods and techniques are being enhanced day by day by the researchers. Review of few techniques and methods used in Big Data analytics are provided in Table 5

Table 5. Big Data analytics techniques and models

No	Technique/ Method	Details
1	A/B Testing	This is called as split testing or bucket testing. In this technique, a control object is compared with many test objects for improvements. This is not being used widely in Big Data
2	Associated Rule Mining	This technique is used for relationships among the huge datasets. This would use many algorithms to generate association rules. Ex. Market basket analysis – it would provide details to a retailer how customers are buying two or more different products together. Which will define an association between those two products. This technique used in data mining
3	Classification	This is a technique to identify the category of the incoming new data based on the existing set which are already categorized based on some data points. This is often used as supervisory learning technique as it uses existing data sets called training dataset. This technique is also used in data mining
4	Cluster Analysis	This is a technique used to group a set of objects into same group based on some commonalities to each other and different to other groups. This technique is used in data mining
5	Crowdsourcing	This is a technique used to generate the ideas/ innovations from a large group through an online method. It creates a new knowledge repository for ideas
6	Data Fusion and Integration	This is a technique used to generate more efficient and potential insights of the data using set of models and techniques which integrate multiple data sets and analyze, rather analyzing them independently
7	Data Mining	This is major technique used for predictive analytics. This enables to identity hidden patterns with help of association rule, classification, regression and cluster analysis on large datasets
8	Ensemble Learning	It is a type of supervisory learning technique. It uses many predictive models such as machine learning and statistical approach to gain better predictions than simple constituent model
9	Genetic Algorithm	This technique is based on the process of Natural evolution

Table 5 (c o n t i n u e d)

No	Technique/ Method	Details
10	Machine Learning	This is a technique used to create artificial intelligence by providing some knowledge to the system. The knowledge is based on the algorithms that enable systems to progress with the help of pragmatic data (knowledge). It is aimed to identify the complex pattern and take appropriative decision based on the intelligence of the system (knowledge)
11	Natural Language Processing	It is a kind of Machine Learning process. NLP uses set of techniques as defined in Machine learning. Ex Sentiment analysis on a social media to find out, how customers are reacting to campaign
12	Neural Networks	It is computational techniques, which was mimic of structure of biological neural networks. It is suitable for nonlinear pattern recognition. It uses both supervised learning and unsupervised learning
13	Network Analysis	This is a technique used in a network or a graph to illustrate the associations among the discrete nodes
14	Predictive Modeling	This is a model built with the help of some of the statistical and mathematical techniques to predict a best possible outcome
15	Regression	This is a statistical technique used for prediction which will determine co-variance between dependent and independent variables
16	Sentiment Analysis	This technique is a Natural Language Processing which can determine the information of subject from the source of textual data
17	Spatial Analysis	This Technique uses set of statistical models which can be used to explore the geographical data such as Geographical Information Systems (GIS) which used to capture the longitude and latitude, i.e., exact location
18	Statistics	Statistics is a science of collecting data in the form of surveys or experiments etc., organizing the data in any sorting order or lexical order etc., with an art of interpretation of the data in a form of hypothesis
19	Supervised Learning	This is a technique which adapts some of the machine learning techniques which can deduce relationships based on prior knowledge (training data)
20	Simulation	This model is used to mimic the behavior of some complex systems to predict and plan the outcome and measure the results
21	Unsupervised Learning	It is a technique uses some of the machine learning models which can identify the hidden patterns from the data without any prior knowledge. An example of Unsupervised learning is Cluster Analysis
22	Visualization	This technique used to provide a gist of the information in very simple way in terms of a graph, diagram, image or any visual representation to simplify the understanding

Since Big Data is a combination of structured, semi-structured and unstructured data sets, it is important to think through various analytical data models for analyzing them. Relational data (structured) can be analyzed with Structured Analytics which is being used commonly to analyse OLAP data. Text analytics are the analytics used for social media analytics, web analytics. Social media analytics is majorly used for sentiment analysis. There are several models available for performing sentiment analysis. There are three types of sentiments [24], positive, negative and neutral, which can bring in a potential business values. Social media data fed by consumer for a product of an organization help the organization to understand the customer sentiments and plan their strategies to improve their sales and product quality. Visual analytics used for analyzing multimedia data such as visual shapes generated by geo special systems, etc., are briefly discussed in the next sections.

4.7. Text analytics

There is an increasing generation of text data in the form of unstructured data such as social media messages, web logs, different forums sharing the feedbacks about some product and email communication, etc. These scenarios have evolved from the traditional structured data usage using RDBMS and Data Warehouses [25] to cater semi structured and unstructured structured data. With digital transformation, IOT, etc., over 90% of data generated is semi-structured or unstructured. So, it is important to use the analytics on these varieties of data. Text analytics is process having multiple steps starting from text data identification to visualization. As shown in Fig. 9, 10 steps involved in text mining, same is described in the next paragraph.

Fig. 9. Text analytics in Big Data

Ten prominent steps involving the Text Analytics (TA) are: 1. Collecting Text Data from the different Sources. 2. Extracting the concepts, entities, events and relations. 3. Creation of taxonomies out of them. 4. Search accessing, web crawling, indexing, de-duplicating from the data clustered and categorized. 5. Analyze the file format using nature language semantics. 6. Build ER modeling. 7. Link the analysis among them. 8. Ability to identify sentiments out of this. 9. Document the summarized information. 10. Visualize them using the tolls to have quick insights

Text Mining (TM) is process to analyze text data to derive possible insights from text data content such as emails, text documents, text communication in the social networking such as postings, feedbacks, etc. Mining unstructured data with Natural Language Processing (NLP), statistical modelling and machine learning techniques is a challenging task, because of natural language text is often inconsistent. Text data may contain some ambiguities due to inconsistent syntax and semantics, including slang, specific business languages, language by different age groups, etc.

4.8. Visual analytics

Visual analytics applications are mostly having huge and high dimensional data sets [26] such as climate research, geo special research and financial market research, etc. Recent developments in visual shape analytics has been a considerable success which requires certain algorithms for analysis for navigating and visualizing that are capable of interactive performance.

The main goal of visual analytic's [27] is to provide insight from very huge datasets such as scientific researches, forensic data, academic records, any business data, HTML/ XML files, Web pages, metadata of any visual database and source code, etc., which are stored in many heterogeneous data stores like NoSQL databases. To bring in the statistical analytics from these data, it is required to iteratively collect and pre-process data [27], however to enable the decision making from these data one has to acquire knowledge and get a perception out of the data by exploring different knowledge representations of the data along with human capacities. Hence, the human factor is a prime factor in the field of visual analytics. The Human Computer Interaction (HCI) [28] is a crucial component that supports knowledge discovery from the visual analytics. HCI is all about the building knowledge of interface between human and computer. The other component in this field is Semi-autonomous system [28], in this model enables to run the system independently for under assured conditions, however this can't take end to end decisions.

4.9. Big Data use cases

There are many tools, techniques, solutions and frameworks are available to implement Big Data and are developed with the references to heterogeneous architectures [29]. Social networking companies such as Linked In, Facebook, Twitter are widely using the Big Data use cases [29].

Table 6. Big Data use cases – industry wide

No	Industry	Use cases
1	Health Care	• Trends of QoS in relevance with health habits[31] • Clinical Decision support system • Remote Patient monitoring • Patient Profiling • Comparative effective research • Disease prediction[32]
2	Marketing	• Cross Selling • Location Based Marketing • Analyzing instore behavior • Sentiment Analytics
3	Manufacturing	• Sensor based operations • Supply chain and inventory management • Shorten driven to value cycle
4	Public Sector/ Govt Sector	• Identifying the basic needs from the public[32] • Traffic decongestion • Civic compliance to reduce Noise, Air and Water pollution
5	Banking & Insurance	• Fraud detection • Customer Predictions • Risk analysis
6	Telecom	• Geo Targeted Ads • Emergency response • Remote monitoring of personal things • Urban planning
7	Oil & Gas	• Remote drilling • Oil Prediction • Digital Oil field

Banking sectors also use many of the use cases, for an instance a bank requires to predict its customer behaviour for successful banking business. A framework called iCARE [30] made an attempt to analyse the unified customer behaviour and their transactions to predict the customer behaviour and giving insights for building new strategies to their business growth. Here is given some of the Big Data use cases used in important industries as detailed in Table 6.

5. Big Data solution

5.1. Apache hadoop – history, architecture and eco system

Hadoop [33] was developed by Dough Cutting along with Mike Cafarella in 2005 [34]. Based on Google's white paper Google's distributed File System (GFS), Hadoop [14] is a distributed and parallel computing paradigm built based on Google's MapReduce paradigm [35]. It is built to process a very huge data sets with in a clustered environment on scale out approach, which means essentially, we can add thousands of machines on to the hadoop cluster, so each offering local computation and storage and can process the data in a parallel.

The Apache Hadoop framework is modules as a master and slave kind of architecture. It contains the following components: 1) Hadoop Distributed File System (HDFS) – this module is the one which has actually store the data in a distrusted fashion; it enables the users by providing access to the data with a high-throughput. 2) Processing the data using (YARN) [35, 36]: Yet another resource negotiator is module provides resource management in the cluster. 3) Hadoop Map Reduce – this module works on top of YARN for processing the large data sets in a parallel processing mode.

Hadoop Distributed File System (HDFS) is the file system in Hadoop framework. It is designed in a distributed approach and to run on commodity hardware [37]. HDFS is highly fault tolerant on low cost hardware. It can store a large amount of data and provides easy access to the users. Master Slave architecture [38] is followed in HDFS. Name node is a master server manages the name spaces in the file system as a meta data [34]. Data Node is a slave system which manages the data storage of the system [34]. In a cluster, there will be multiple data nodes which can be used for data operations. Block Size is the minimum amount of data is being read or write by HDFS file system at once. The size of the block is generally 64 MB [34].

MapReduce [39] is a framework used for processing the batch processing jobs in the HDFS work on parallel processing mode in the cluster [40] and it is programming model [38]. The program breaks the execution into two different parts or phrases such as Map phase and Reducer Phase [34]. *YARN* - Yet Another Resource Negotiator is cluster management system in the Apache Hadoop later versions [39]

Some related projects/tools in Hadoop ecosystem are depicted in the following Fig. 10.

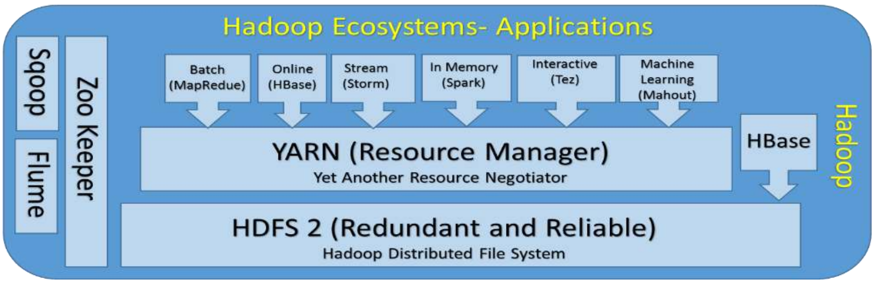

Fig. 10. Hadoop technology stack

Apache Hive [39] is a data warehousing package[5] that is built on top of Hadoop [41] for creating database, tables/views, etc. This is mainly used for data query and analytics [34]. It is very similar to SQL, which is called HiveQL [42, 43].

Apache Pig [39] an open source high level data flow system for analysing huge data sets due to amendable structure it enables greater parallelism [44]. This tool is a high level query and analysis tool built to overcome the obstacles of writing map reduce code for non-java programmers Pig was created [5]. The language is called Pig Latin [34]. This will be compiled into map reduce jobs by Pig compiler that will run on Hadoop [45].

Apache Flume is a data acquisition tool for data integration [46], which is distributed, highly reliable and with simple configurations and primarily for streaming data processing such as log data from various web servers to HDFS [34].

Apache Sqoop is a data acquisition tool designed for efficiently transferring the bulk data between hadoop and structured data stores such as relational database [47].

Apache Spark [36] is a lightning-fast cluster computing framework designed for fast computation. It can run on standalone, hadoop, mesos or even cloud [48]. Spark is lightning faster because of the fact it is built in-memory computing framework to address one of the original hadoops disadvantage iterative operations [49]. In memory cluster processing and cluster computing is the prime feature of Spark [50], hence it can have greater processing speed. It covers batch, streaming, interactive and iterative workloads [51]. Resilient Distributed Datasets (RDDs) is core data units in Spark [50]. Spark is becoming popular as it has features such as speed, multi-language support and analytics support.

Apache Zookeeper is a software project from Apache, providing an open source distributed configuration service, synchronization service and naming registry for large distributed systems [52]. It will manage and co-ordinate clusters (like Hbase, Hadoop, Solr, etc.).

As mentioned, there are many advantages of Hadoop usage for Big Data challenges and few of them are given in brief:

1. Distributed data storage and processing framework work with computation done at local to data which can prevents the network overload.
2. Linear scale out compared to RDBMS or any other solution.
3. Fault tolerant and high availability.
4. Ability to process in parallel and very rapidly of large data sets.
5. Has ability to process wide varieties of data and stream of data.
6. It best suits for data ware housing for loaded data into Hadoop [53].

Some of the disadvantages of distributed computing using Hadoop:

1. Map reduce programming model is very restrictive due to lack of central data can be preventive as it is not meant for iterative and recursion.

2. Joins of multiple datasets are tricky and slow hence there is often entire dataset gets copied in the process, which means essentially de normalize the data for better performance.

3. It will not work well with small and many data sets.

6. Conclusions and further scope

This review article we have discussed overview of data transformed to Big Data, challenges and issues in Big Data processing and how to harness the data using the Big Data analytics and models. We have featured some of the solutions for Big Data challenges with technology advancements and their implementations in today's world, being used by industries to provide state of art and near real time data analytics solution with high precision enhancing the ability to understand the end user experience and provide better business models.

However, there are some open challenges with these Big Data solutions such as: 1) fast changing system configuration requirements due to highly dynamic workload constraints, varying innovation cycles of system hardware components, 2) low latency querying, 3) transactional data handling, 4) data normalization is not feasible etc. which are not addressed in the current solution. There is an interesting proposal by H.-K. Lin and others, an architectural system called composable systems which can potentially address this challenge by rack scale architecture [53]. Secondly, there is a need of next gen Big Data solutions which could take advantages for transactional management in a scaling out model like other, not only SQL to bridge the gap. There are experiments in the similar lines done by D. Plase and team in order to gain more detailed experience with compact data formats [54] and an interesting architecture called Splice Machine [55]. Splice Machine is coming up with a hybrid approach to take an advantage to support greater concurrency even for small data writes and reads on a distributed scaled out technologies. Apart from this, other major challenge is analysing Big Data at real time. The current technologies and frameworks are not in a position to handle all the real-time scenarios to analyse the data at real time. There is an attempt made by K. Wang and others for real time analytics using hybrid-stream a Big Data analytics model [56]. There is an interesting novel proposal for normalizing the data in Big Data approach by Golov and Rönnbäck [57] using Anchor modelling technique. This seems to have efficient way to store and process the data. There is a potential scope for research in the above-mentioned challenges.

References

1. Demchenko, Y., C. D. Laat, P. Membrey. Defining Architecture Components of the Big Data Ecosystem. – In: Proc. of International Conference Collaboration Technologies and Systems (CTS'14), Vol. **14**, 2014, pp. 104-112.

2. S l a v a k i s, K., G. B. G i a n n a k i s, G. M a t e o s. Modeling and Optimization for Big Data Analytics: (Statistical) Learning Tools for Our Era of Data Deluge. – IEEE Signal Processing Magazine, Vol. **31**, 2014, pp. 18-31.
3. S h e r m a n, R. Chapter 1 – The Business Demand for Data, Information, and Analytics. – Business Intelligence Guidebook, Morgan Kaufmann, Boston, 2015, pp. 3-19.
4. L i n s t e d t, D., M. O l s c h i m k e. Chapter 1 – Introduction to Data Warehousing – In Data Vault 2.0, Morgan Kaufmann, Boston, 2016, pp. 1-15.
5. S h a r m a, S. Expanded Cloud Plumes Hiding Big Data Ecosystem. – Future Generation Computer Systems, Vol. **59**, 2016, pp. 63-92.
6. C o h e n, J., B. D o l a n, M. D u n l a p, J. M. H e l l e r s t e i n, C. W e l t o n. MAD Skills: New Analysis Practices for Big Data. – Proc. VLDB Endow, Vol. **2**, 2009, pp. 1481-1492.
7. H u, H., Y. W e n, T. S. C h u a, X. L i. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. – IEEE Access, Vol. **2**, 2014, pp. 652-687.
8. M y e r s o n, J. M. Cloud Computing Versus Grid Computing. 3 March 2009. **http://www.ibm.com/developerworks/library/wa-cloudgrid/**
9. A l k h a n a k, E. N., S. P. L e e, R. R e z a e i, R. M. P a r i z i. Cost Optimization Approaches for Scientific Workflow Scheduling in Cloud and Grid Computing: A Review, Classifications, and Open Issues. – Journal of Systems and Software, Vol. **113**, 2016, pp. 1-26.
10. The Digital Universe of Opportunities: Rich Data Increasing Value of the Internet of Things. – EMC Digital Universe with Research & Analysis by IDC. **http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm**
11. K i m, L. Here's What Happens in 60 Seconds on the Internet. 11 December 2015. **http://smallbiztrends.com/2015/12/60-seconds-on-the-internet.html**
12. K a r t, N. H. L., F. B u y t e n d i j k. Survey Analysis: Big Data Adoption in 2013 Shows Substance behind the Hype. – Gartner's 2013 Big Data Study, 2013.
13. Contributors, W. Big Data. 12 March 2016. UTC. **https://en.wikipedia.org/w/index.php?title=Big_data&oldid=709642525**
14. I s h w a r a p p a, J. A n u r a d h a. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. – Procedia Computer Science, Vol. **48**, 2015, pp. 319-324.
15. W a t s o n, H. J. Tutorial: Big Data Analytics: Concepts, Technology, and Applications. – Association for Informaiton Systems, Vol. **34**, 2014, pp. 5-16.
16. S w a n, M. Philosophy of Big Data: Expanding the Human-Data Relation with Big Data Science Services. – In: Proc. of First International IEEE Conference of Big Data Computing Service and Applications (BigDataService'2015), 2015, pp. 468-477.
17. F a r i d, M., A. R o a t i s, I. F. I l y a s, H.-F. H o f f m a n n, X. C h u. CLAMS: Bringing Quality to Data Lakes. – In: Proc. of 2016 International Conference on Management of Data, San Francisco, California, USA, 2016, pp. 2089-2092.
18. D o n K o g a n. Top 8 Bigdata Trends 2016. – White Paper, 2016.
19. R i t h, J., P. S. L e h m a y r, K. M e y e r-W e g e n e r. Speaking in Tongues: SQL Access to NoSQL Systems. – In: Proc. of 29th Annual ACM Symposium on Applied Computing, Gyeongju, Republic of Korea, 2014, pp. 855-857.
20. G a i t h o, M. How Applications of Big Data Drive Industries. – Simplylearn. **http://www.simplilearn.com/big-data-applications-in-industries-article**
21. S h e r m a n, R. Chapter 15. Advanced Analytics. – In: Business Intelligence Guidebook. Boston, Morgan Kaufmann, 2015, pp. 375-402.
22. G a n d o m i, A., M. H a i d e r. Beyond the Hype: Big Data Concepts, Methods, and Analytics. – International Journal of Information Management, Vol. **35**, 2015, pp. 137-144.
23. M a n y i k a, M. C. J., B. B r o w n, J. B u g h i n, R. D o b b s, C. R o x b u r g h, A. H. B y e r s. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, June 2011.
24. V a t r a p u, R., R. R. M u k k a m a l a, A. H u s s a i n, B. F l e s c h. Social Set Analysis: A Set Theoretical Approach to Big Data Analytics. – IEEE Access, Vol. **4**, 2016, pp. 2542-2571.
25. I t t o o, A., L. M. N g u y e n, A. V a n D e n B o s c h. Text Analytics in Industry: Challenges, Desiderata and Trends. – Computers in Industry, Vol. **78**, 2016, pp. 96-107.
26. H e r m a n n, M., R. K l e i n. A Visual Analytics Perspective on Shape Analysis: State of the Art and Future Prospects. – Computers & Graphics, Vol. **53**, Part A, 2015, pp. 63-71.

27. González-Torres, A., F. J. García-Peñalvo, R. Therón-Sánchez, R. Colomo-Palacios. Knowledge Discovery in Software Teams by Means of Evolutionary Visual Software Analytics. – Science of Computer Programming, Vol. **121**, 2016, pp. 55-74.
28. Makonin, S., D. McVeigh, W. Stuerzlinger, K. Tran, F. Popowich. Mixed-Initiative for Big Data: The Intersection of Human + Visual Analytics + Prediction. – In: 2016 49th Hawaii International Conference on System Sciences (HICSS'16), 2016, pp. 1427-1436.
29. Pääkkönen, P., D. Pakkala. Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems. – Big DATA Research, Vol. **2**, 2015, pp. 166-186.
30. Sun, N., J. G. Morris, J. Xu, X. Zhu, M. Xie. iCARE: A Framework for Big Data-Based Banking Customer Analytics. – IBM Journal of Research and Development, Vol. **58**, 2014, pp. 4:1-4:9.
31. Batarseh, F. A., E. A. Latif. Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare. – Big Data Research, Vol. **4**, 2016, pp. 13-24.
32. Archenaa, J., E. A. M. Anita. A Survey of Big Data Analytics in Healthcare and Government. – Procedia Computer Science, Vol. **50**, 2015, pp. 408-413.
33. Saraladevi, B., N. Pazhaniraja, P. V. Paul, M. S. S. Basha, P. Dhavachelvan. Big Data and Hadoop – a Study in Security Perspective. – Procedia Computer Science, Vol. **50**, 2015, pp. 596-601.
34. Uzunkaya, C., T. Ensari, Y. Kavurucu. Hadoop Ecosystem and Its Analysis on Tweets. – Procedia – Social and Behavioral Sciences, Vol. **195**, 2015, pp. 1890-1897.
35. Cassales, G. W., A. S. Charão, M. K. Pinheiro, C. Souveyet, L. A. Steffenel. Context-Aware Scheduling for Apache Hadoop over Pervasive Environments. – Procedia Computer Science, Vol. **52**, 2015, pp. 202-209.
36. Shyam, R., B. H. B. Ganesh, S. S. Kumar, P. Poornachandran, K. P. Soman. Apache Spark a Big Data Analytics Platform for Smart Grid. – Procedia Technology, Vol. **21**, 2015, pp. 171-178.
37. Ma, Y., Y. Zhou, Y. Yu, C. Peng, Z. Wang, S. Du. A Novel Approach for Improving Security and Storage Efficiency on HDFS. – Procedia Computer Science, Vol. **52**, 2015, pp. 631-635.
38. Maitrey, S., C. K. Jha. MapReduce: Simplified Data Analysis of Big Data. – Procedia Computer Science, Vol. **57**, 2015, pp. 563-571.
39. Loshin, D. Chapter 7. Big Data Tools and Techniques. – In: Big Data Analytics. Boston, Morgan Kaufmann, 2013, pp. 61-72.
40. Yildiz, O., S. Ibrahim, G. Antoniu. Enabling Fast Failure Recovery in Shared Hadoop Clusters: Towards Failure-Aware Scheduling. – Future Generation Computer Systems, 2016.
41. Apache Hive TM.
https://hive.apache.org/
42. Chennamsetty, H., S. Chalasani, D. Riley. Predictive Analytics on Electronic Health Records (EHRs) Using Hadoop and Hive. – In: 2015 IEEE International Conference Electrical, Computer and Communication Technologies (ICECCT'15), 2015, pp. 1-5.
43. Xu, Y., S. Hu. QMapper: A Tool for SQL Optimization on Hive Using Query Rewriting. – In: Proc. of 22nd International Conference on World Wide Web, Rio De Janeiro, Brazil, ACM, Vol. **1**, 2013, pp. 211-212.
44. Apache Pig.
https://pig.apache.org/
45. Rajurkar, G. D., R. M. Goudar. Notice of Violation of IEEE Publication Principles, A Speedy Data Uploading Approach for Twitter Trend and Sentiment Analysis Using HADOOP. – In: International Conference on Computing Communication Control and Automation (ICCUBEA'15), Vol. **1**, 2015, pp. 580-584.
46. Apache Flume.
https://flume.apache.org/
47. Apache Sqoop.
http://sqoop.apache.org/
48. Apache Spark.
http://spark.apache.org/

49. L i, H., K. L u, S. M e n g. Bigprovision: A Provisioning Framework for Big Data Analytics. – IEEE Network, Vol. **29**, 2015, pp. 50-56.
50. R e y e s-O r t i z, J. L., L. O n e t o, D. A n g u i t a. Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf. – Procedia Computer Science, Vol. **53**, 2015, pp. 121-130.
51. E l i a, D., S. F i o r e, A. D'A n c a, C. P a l a z z o, I. F o s t e r, D. N. W i l l i a m s. An In-Memory Based Framework for Scientific Data Analytics. – In: Proc. of ACM International Conference on Computing Frontiers, 2016, pp. 424-429.
52. Apache ZooKeeper™.
https://zookeeper.apache.org/
53. L i n, H.-K., J. A. H a r d i n g, C.-I. C h e n. A Hyperconnected Manufacturing Collaboration System Using the Semantic Web and Hadoop Ecosystem System. – Procedia CIRP, Vol. **52**, 2016, pp. 18-23.
54. P l a s e, D., L. N i e d r i t e, R. T a r a n o v s. Accelerating Data Queries on Hadoop Framework by Using Compact Data Formats. – In: 4th IEEE Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE'16), 2016, pp. 1-7.
55. Splice Machine.
http://www.splicemachine.com/product/
56. W a n g, K., J. M i, C. X u, L. S h u, D. J. D e n g. Real-Time Big Data Analytics for Multimedia Transmission and Storage. – In: IEEE/CIC International Conference on Communications in China (ICCC'16), 2016, pp. 1-6.
57. G o l o v, N., L. R ö n n b ä c k. Big Data Normalization for Massively Parallel Processing Databases. Computer Standards & Interfaces Available Online, 2017. ISSN 0920-5489.